Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment

David Masclet, Laurent Denant-Boèmont, Charles Noussair

Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment

Laurent Denant-Boemont, David Masclet and Charles Noussair[1]

July 2005

Abstract

We present the results of an experiment that explores the sanctioning behavior of individuals who experience a social dilemma. In the game we study, players choose contribution levels to a public good and subsequently have multiple opportunities to reduce the earnings of the other members of the group. The treatments vary in terms of individuals' opportunities to (a) avenge sanctions that have been directed toward themselves, and (b) punish others' sanctioning behavior with respect to third parties. We find that the individuals avenge sanctions they have received, punish those who fail to sanction third parties, and punish low contributors, even when punishment is costly to the sanctioner. When there are five rounds of unrestricted sanctioning, contributions and welfare are significantly lower than when only one round of sanctioning opportunities exists, and welfare is lower than the zero-cooperation benchmark.

1. Introduction

One focus of experimental research on social dilemmas has been the search for and the identification of factors that promote cooperative behavior in settings in which individuals have incentives to behave opportunistically. The most widely used arena for this investigation is the voluntary contributions mechanism, an elegant construction that permits straightforward measurement of the extent of self-versus group-interested behavior. Interaction in the voluntary contributions mechanism proceeds according to the following rules. Each member of a group of individuals has an endowment, from which he may contribute any amount to a public good that returns a payoff to each individual. The level of this payoff ensures that at the social optimum, each individual contributes his entire endowment while, in contrast, each individual has a dominant strategy to contribute zero. The amount contributed can be interpreted as a measure of cooperative behavior. The main overall pattern observed in laboratory experiments is that initial contributions are substantial, but decline as the game is repeated and cooperation converges to a

[1]Denant-Boemont: Université Rennes 1, Rennes, France. Masclet: Université Rennes 1, Rennes, France. Correspondence to Noussair: Department of Economics, Emory University, Atlanta, GA 30322, USA. Tel: 1-404-712-8167. Fax: 1-404-727-4639. E-mail: cnoussa@emory.edu. We thank participants in seminars at the University of Wisconsin-Madison, the 2004 ESA Meetings in Tucson, Arizona, USA, the 2004 IMEBE Meetings in Cordoba, Spain, and the 2005 SAET meetings in Vigo, Spain, for constructive and helpful comments. We thank Elven Priour for programming and organization of the sessions.

near-negligible level in the long run (Isaac et al., 1984; Andreoni, 1988; Isaac and Walker, 1988a; Ledyard, 1995).

However, a number of modifications to the game that increase cooperation considerably, even in the long run, have been identified.[2] For example, endowing individuals with the ability to reduce the earnings of the least cooperative individuals in the group is highly effective in raising contribution levels (see for example Yamagishi (1986), Ostrom et al (1992), Fehr and Gaechter (2000), Carpenter (2004), Bochet et al. (2005), and Masclet et al. (2003)). These studies all find that individuals are willing to pay from their own earnings to reduce the earnings of free riders, and average contributions increase as a result of the existence of the sanctioning opportunity. It is thus clear that, at least under some circumstances, sanctioning mechanisms can represent an effective means of increasing cooperation among individuals and thus alleviate free-rider problems. This is the case even when the punishment is costly for sanctioners to administer,[3] and when the system does not rely on an external trigger mechanism for enforcement.

Immunity of sanctioners from reprisals is a characteristic of all of the studies listed in the last paragraph. Because there is only one opportunity to sanction in each period, and there is no means to track the identity of others from period to period, no player can identify individual punishment behavior in a manner that allows him to target an individual for reciprocation.[4] If such reprisals were possible, it might deter sanctioning, and thus dilute the effectiveness of the system in increasing contribution levels. Nikiforakis (2004) reports an experiment focused on this issue. He conducts an experiment in which agents may punish those, and only those, who punished them in the current period[5]. Nikiforakis terms the avenging of sanctions as "counterpunishment", and he finds that counterpunishment nearly offsets the increase in contributions the existence of the opportunity to punish creates.

On the other hand, all of the above studies also preclude the use of punishment to enforce the sanctioning regime. Because individuals who administer sanctions bear the cost of doing so, while all players benefit from the resulting increase in contributions, there is an incentive for individuals to free ride on others' provision of sanctions against low contributors. However, there is no mechanism for targeting

---

[2] These include preplay communication (Isaac and Walker, 1988b), creation of group identification in conjunction with post-play open discussion (Gaechter and Fehr, 1999), and having each individual assign a rating to each other group member's contribution decisions (Masclet et al., 2003).

[3] See Falk et al (2005) for a detailed analysis of the motivation behind the application of costly sanctions.

[4] In some of the studies, in which group membership was fixed, agents could punish all other group members by contributing less or by randomly sanctioning other agents in the following rounds, but an individual could not be targeted for sanctions based on his prior sanctioning behavior.

[5] Nikiforakis' experiment is related to ours in that in there are two rounds of sanctions. However, his design differs from ours in that the agents become aware after the first round of sanctions only of the quantity of sanctions that each individual assigned to him, and only have the opportunity to sanction those who sanctioned him. This allows sanctions in the second round only for the purpose of counter-punishment (avenging sanctions received), and not for the purposes of punishing those who did not sanction free riders or as a second opportunity to punish low contributors. Thus, his design may be particularly conducive to a reduction in contribution levels.

punishment toward the individuals who free ride on others' sanctions. Thus, it may be the case that the possibility of reprisals against those who fail to apply sufficient sanctions may induce more sanctioning of low contributors, and thereby increase contributions. We will refer to reprisals against low sanctioners as "sanction enforcement".[6]

In this paper, we consider whether the effective self-governance that sanctioning opportunities create is robust to the introduction of the ability to observe and punish all sanctioning decisions of other players. The key to answering this question is the relative strength of the effects of counterpunishment and sanction enforcement on contributions. Studying the effect of additional rounds of sanctioning is of empirical relevance as imperfect anonymity for contribution and punishment behavior is a characteristic of many social dilemmas in the field, such as pollution reduction, industry cartels, and team projects. The design of our experiment allows us to study in detail patterns of counterpunishment, sanction enforcement, and punishment for low contributions, and the resulting effects on subsequent contribution and punishment levels, for a game in which there are two rounds of sanctions. The existence of exactly two stages allows isolation of the effects of counterpunishment and sanction enforcement on contributions.

As discussed in detail in section three, we find that counterpunishment exerts a downward effect of contributions that the increase in contributions from sanction enforcement does not fully offset. We then explore whether the effects are magnified when there are a greater number of, namely five, stages of sanctions. There we find that contributions and welfare levels are significantly lower than when there is only one round. Indeed, *welfare levels are lower than the minimum levels possible in the absence of a sanctioning mechanism*. The details of our experimental design are presented in section two. We present the results of the experiment in section three and offer some concluding thoughts in section four.

---

[6] Two recent studies suggest that the tendency to enforce sanctions may be weak. Kiyonari and Barclay (2005) allow individuals a second round of sanctions. This allows individuals to sanction other players (who were computerized unbeknownst to participants) based on these other players' punishment behavior in the first round. The authors find little evidence of a tendency to sanction those who sanctioned relatively low amounts during the first stage. Because their experiment consists of a one-shot game, punishment cannot affect recipient's future behavior. Cinyabuguma et al. (2004) report an experiment in which a second sanctioning opportunity is available after every third period. Contrary to the pattern one would expect to observe under sanction enforcement, individuals who punish more during the first punishment opportunity receive more punishment in the second. However, it is unclear how much of this punishment of high sanctioners in their study consists of counterpunishment. Subjects in their study observe how much of other players' punishment assignments are directed toward above-average, below-average, and average contributors, and can condition their second round of punishment on this information. Those engaging in "perverse punishment", the punishment of high contributors, are targeted most severely during the second opportunity to sanction. The authors find that the treatment, in which a second opportunity to sanction is available, is characterized by higher contributions and earnings compared to a situation where no second punishment opportunity exists. They find that the existence of the second punishment opportunity reduces sanctioning in the first punishment stage, and indeed the total amount of punishment in the two stages combined is less than when there is only one stage. Cinyabuguma et al. also find that the receipt of sanctions in the second sanctioning phase reduces sanctioning in the first phase of the subsequent period. By analyzing data from the last period of their sessions in which a second stage

2. The Experiment

2.1 Overview

There are five treatments in the experiment, all of which have a first and a second stage of interaction in common. In the first stage of the game, players simultaneously decide how much of their endowment to contribute to the public good. In the second stage, players are informed of the decisions that other members of their group have made and have the opportunity to punish them. The punishment reduces the earnings of both the sanctioning and the sanctioned parties. In the *Benchmark* treatment, the two stages described above comprise all of the activity in the game. The benchmark treatment is a replication of Fehr and Gaechter (2000). In the other treatments, in contrast, there are additional stages, in which players are informed about some or all of the sanctioning activity in stage two, and any individual may sanction any or all members of their group in a manner similar to stage two.

Three of the new treatments consist of exactly three stages, one contribution and two sanctioning stages, and differ from each other in terms of the information available to players about the identity of those administering sanctions in the second stage. In the *Full Information* treatment, players are informed about how much each individual sanctioned each other individual. In the *Revenge Only* treatment, each individual is informed of the source and the quantity of the sanctions directed toward him, but does not know how much other members of the group were sanctioned and by whom. In the *No Revenge* treatment, players are not informed of the origin and the distribution of the sanctions they receive, but do learn who sanctioned other group members and by how much. The information received at the end of the second stage may be used in the determination of punishment assignment strategies in the third stage.

The Benchmark treatment allows punishment only in response to contribution decisions. Previous research indicates that most of this punishment is directed at relatively low contributors, although sanctioning of high contributors is also observed (Cinyabuguma et al., 2004), although its incidence varies depending on the subject pool employed (Gaechter and Herrmann, 2004). The Revenge Only treatment allows counterpunishment as well as punishment for contribution behavior in the first stage. The No Revenge treatment allows sanction enforcement and punishment for low contributions, but precludes counterpunishment. The Full Information treatment allows sanction enforcement, counterpunishment, and punishment for low contributions. Therefore, the difference in contributions between the Benchmark and the Revenge Only treatments, as well as the difference between the No Revenge and the Full Information treatments, measure the effect of counterpunishment on contributions. The first and second comparisons

---

of punishment is possible, they note that punishment of perverse punishers is less likely to be strategic than punishment of others.

are within an environment in which sanction enforcement is impossible and possible, respectively. We hypothesize, based on the results of Nikiforakis (2004), that the effect of the possibility of counterpunishment on contributions is negative.

The difference in contributions between the Benchmark and the No Revenge treatment, as well as the difference between the Revenge Only and the Full Information treatment, is interpreted as the effect of the introduction of sanction enforcement. We hypothesize that the effect of sanction enforcement on contributions is positive. The difference between the Full Information and the Benchmark treatments measures the effect of removing all immunity from reprisals for sanctioning decisions, incorporating the effects of both counterpunishment and sanction enforcement, as well as the opportunity to delay and update punishment for low contributions. Table 1 below summarizes the anticipated effect of each of the motives for punishment on contributions and serves as the basis for the following statement summarizing our hypotheses concerning treatment differences.

$$C(NR) > C(B) = C(FI) > C(RO) \qquad (1)$$

Where C(X) refers to the average amount contributed in treatment X, and NR, B, FI, and RO are abbreviations for the four treatments. In table 1, a (+) indicates the existence of a hypothesized positive effect on contributions, and a (-) denotes a negative effect. Because No Revenge allows two positive effects on contribution levels, sanction enforcement and punishment for low contributions, and allows no counterpunishment, we anticipate that it would have the highest contribution levels. B differs from NR in that no sanction enforcement is possible so that only sanctions of low contributions can promote high contributions. FI differs from NR in that it adds the potential negative effect of counterpunishment. FI differs from B in that both sanction enforcement and counterpunishment are added, and we hypothesize a priori that the effect of the two motives on contributions offset so that average contributions are equal in the two treatments. Finally, because RO differs from FI only in that it precludes sanction enforcement and from B only in that it permits counterpunishment, we hypothesize that it generates lower contribution levels than either FI or B.

[Table 1: About Here]

The final treatment is the Six-Stage Full Information treatment, 6SFI. In this treatment, as in the other four, the first stage of the game consists of players simultaneously choosing how much of their endowment to contribute to the public good. In the second stage, players are informed of the decisions that other members of their group have made and have the opportunity to punish them, as in the Full

Information treatment. Stages 3 – 6 are identical to stage 2. Players observe the sanctioning decisions of all members of the group in earlier stages and then may sanction any other individuals. The Six Stage Full Information treatment allows punishment for complex motivations such as punishing counterpunishment or punishing a failure to enforce sanctions. However, we will maintain the hypothesis a priori, that the behavior that the additional motives to punish that the extra stages introduce do not cause a consistent change in average contributions, and therefore that:

$$C(B) = C(FI) = C(6SFI) \qquad (2)$$

2.2 Procedures

The experiment consists of ten sessions, two sessions conducted under each of the five treatments. An average of twelve individuals participated in each session, for a total of 120 participants.[7] All sessions were conducted at the LABEX of the University of Rennes I, Rennes, France in May of 2004. The experiment was computerized and the scripts were programmed using the z-tree platform (Fischbacher, 1999). The subjects were undergraduate students from a variety of majors. Roughly one-third were economics students in the first two years of college, and all but a small number of the remaining two thirds were students in law, management, and medicine. No individual participated in more than one session.

In each session, there are twenty periods of interaction. Each period within a session proceeds under identical rules. The subjects participating in the session are assigned to groups of size four with fixed membership, in such a manner that they do not know the identities of the other members of their group. The three independent groups in each session, and two sessions per treatment yield six independent observations in each treatment. At the end of each period, individuals remain in the same group. However, individuals' designated labels and the location of the display of their data on the computer screen are reassigned on a random basis in each period. For example, if a player is designated as player *A* in period *t*, he has exactly a ¼ chance of being player *A* in period *t*+1, as well as a ¼ chance of being player *B*, *C*, or *D* in period *t*+1.

The design and the parametric structure of the experiment draw heavily on those of Fehr and Gaechter (2000). At the beginning of each period in all treatments, each participant receives an endowment of 20 ECUs (experimental currency units, 1 ECU = 2 Eurocents). He then must choose to allocate the endowment between a private account, which is his to keep, and a public account, which

[7] In eight of the ten sessions, there were exactly 12 participants. In one of the Six Stage Full Information sessions there were eight participants (comprising two groups) and in the other session there were sixteen participants and thus four groups. Overall, there were six groups, and thus six independent observations, in each of the five treatments.

yields .4 ECUs to each member of the group for each ECU allocated to the account. Following previous authors, we will refer the amount that the individual allocates to the group account as his *contribution*, because the more he allocates to the group account, the lower his own but the greater the group's total earnings. At the end of the first stage, each individual's provisional earnings are equal to $\pi_i^1 = (20 - c_i) + 0.4\sum_{j=1}^{4} c_j$, where c_i is the contribution of player i.

After contribution choices have been made, they are revealed to all group members, and the game enters stage two. Each member of the group is informed of the total contribution of the group and the individual contribution of the three other group members to the public good, as well as her own provisional earnings in the 1st stage, π_i^1. In stage two, players have an opportunity to assign sanctions to each of the other members of their group. Sanctions take the form of an assignment of a number of punishment points in the range from 0 to 10 inclusive. A different number of points and thus a different sanction level may be assigned to each other player. Each point received reduces the recipient's earnings by 10 percent of his first period's earnings from the provision of the public good and his uncontributed endowment. Assigning points is also costly to the sanctioner. The cost function for punishment for player i, denoted as $k_i(p_i^{jm})$, where p_i^{jm} is the number of points that player i assigns to j in stage m,[8] is shown in table 2.

[Table 2: About Here]

The cost function for punishment points is additive across recipients. That is, $k_i(\Sigma_j p_i^{j2}) = \Sigma_j k_i(p_i^{j2})$, The cost function is common to all individuals, so that $k_i(p_i^{j2}) = k_n(p_n^{j2}) = k(p_i^{j2})$. An individual can receive punishment points from any member of the group. Player i's provisional earnings after the second stage are given by:

$$\pi_i^2 = \pi_i^1\left[\max\{0,1-(1/10)\sum_{j\neq i} p_j^{i2})\}\right] - \sum_{j\neq i} k(p_i^{j2}) \qquad (3)$$

8 Suppose for example that player 1 assigns two points to player 2, 9 points to player 3, and 0 points to player 4. Then the total cost to player 1 of the points is 2 + 25 + 0 = 27 ECUs. The 27 ECUs are then reduced from player 1's first stage earnings. Player 2's earnings are reduced by 20% (not including any reductions from points received from other players) and player 3's earnings are reduced by 90%.

Each point the agent receives reduces his earnings by 10% of stage 1 earnings, with a maximum reduction of 100% (receiving more than 10 points imposes no further reduction in earnings, but nonetheless is costly to sanctioners). The cost of punishment is then subtracted to calculate stage two earnings. The two stages described above comprise all of the activity in a period of the baseline treatment.

In the four other treatments, there are subsequent stages of activity. The FI, NR, and RO treatments consist of three stages while 6SFI consists of six stages. The FI, NR, and RO treatments differ only in the information available to each individual after stage two. In the Full Information treatment, each player *i* is informed of the amount that each player sanctioned each other player. That is, he observes p_k^{j2}, for all *j* and *k*. In the No Revenge treatment, player *i* is informed only about how other individuals were sanctioned. That is, player *i* observes p_k^{j2}, for all *k* and for all $j \neq i$, but not for $j = i$. In the Revenge Only treatment, each player is only informed about his own sanctions received. In other words, individual *i* observes p_k^{i2} for all *k*, but does not observe p_k^{j2} for $j \neq i$. The cost of the sanctions to both the sanctioning and the sanctioned parties is identical to stage two in the three treatments. That is $k(p_i^{j3}) = k(p_i^{j2})$. In the three treatments, subjects observe the total number of points assigned to them ($\sum_j p_j^{i2}$ as well as $\sum_j p_j^{i3}$) in each of the two punishment stages.

After the appropriate sanctioning information is transmitted to participants, each member of the group has a second opportunity to assign punishment points to the other players. During this stage, each individual has each other individual's contribution decision available. The cost function for punishment is identical in the third stage (second punishment opportunity) to the second stage (first punishment opportunity), and additive across rounds and recipients, so that $k(p_i^{j2}) = k(p_i^{j3})$ and $k(p_i^{j3}) = k(p_i^{m3})$ for all i and j. As in the earlier punishment stage, each point received reduces an individual's earnings by 10% of her first stage earnings. The final earnings for individual *i* in a period are equal to:

$$\pi_i^3 = \pi_i^1 \left[\max\{0, 1 - (1/10)[\sum_{j \neq i} p_j^{i2} + \sum_{j \neq i} p_j^{i3}]\} \right] - \sum_{j \neq i} k(p_i^{j2}) - \sum_{j \neq i} k(p_i^{j3}) \qquad (4)$$

In the Six Stage Full Information (6SFI) treatment, there are three more stages that follow stage three. These are identical to stages two and three of the FI treatment in that at the end of stage *s*, subject *i* observes $p_j^{k,s-1}$ for all *j*, *k*. The payoff function at the end of stage 6 for individual *i* equals:

$$\pi_i^6 = \pi_i^1 \left[\max\{0, 1 - (1/10) \sum_{s=2}^{6} \sum_{j \neq i} p_j^{is}\} \right] - \sum_{s=2}^{6} \sum_{j \neq i} k(p_i^{js}) \qquad (5)$$

The cost of punishment in 6SFI is additively separable across all stages and individuals, as it is in the other treatments. At the end of each period in all treatments, each participant's computer displays the

contribution of each individual, the sanctioning information about others permitted under the treatment condition, the total quantity of punishment points the individual received in each stage, and his period and accumulated earnings for the session.

The subgame perfect equilibrium of a one-shot version of the game is unique in each treatment. Consider first the Full Information, No Revenge, and the Revenge Only treatments. In stage three there is no punishment in any subgame perfect equilibrium. Because punishment is costly to administer, it is optimal for each player i to set $p_i^{m3} = 0$ at all stage three decision nodes for all other players m. In stage two, similar logic applies. Because all players' stage three equilibrium actions are independent of prior activity, there can be no benefit to assigning sanctions in stage two, and thus it is optimal for each player i to choose $p_i^{m2} = 0$ at all stage two decision nodes for all other players m. Since activity in subsequent stages is independent of decisions in stage 1, it is optimal for all individuals to contribute 0 ECU to the public good. Thus the unique subgame perfect equilibrium of the game is the following. In stage 1, all players contribute zero. In stage 2, all players assign zero punishment to each player regardless of the actions chosen in stage 1. In stage 3, all players assign zero punishment to each other player regardless of actions chosen in stages 1 and 2. In the Baseline and 6SFI treatments, similar logic can be used to show that in any subgame perfect equilibrium, no players punish in any stage, and all players contribute zero in stage 1. If the number of times the game is repeated is common knowledge, and that number is finite, as in our experiment, the above behavior in each repetition of the game constitutes the unique subgame perfect equilibrium to the repeated game.

3. Results

This section is organized as follows. Subsection 3.1 considers patterns in average contributions and earnings in the FI, NR, and RO treatments, the treatments with two punishment stages. The treatments are analyzed in relation to each other and to the Baseline treatment, and the hypotheses advanced in section two are evaluated. Subsections 3.2 and 3.3 study patterns in individual sanctioning behavior in stages two and three, respectively, of the three treatments. Subsections 3.4 and 3.5 consider the effects of the receipt of sanctions in stages two and three respectively, on subsequent behavior. In section 3.6, we report the results for the 6SFI treatment.

3.1 The Effect of a Second Punishment Stage on Contribution and Welfare Levels

Figure 1 illustrates the time path of individual contributions by period, averaged across groups, in the five treatments. The period number is shown on the horizontal axis and the average contribution on the vertical axis. The maximum possible individual contribution, corresponding to the group optimum, is 20.

The minimum possible contribution, corresponding to the Subgame Perfect Equilibrium level, is 0. Figure 2 shows the corresponding time series of average group earnings by treatment. The maximum possible group earnings level, associated with all players contributing their entire endowment and no sanctioning, is 128, while group earnings equal 80 in the Subgame Perfect Equilibrium.[9] The average contribution for each group in each treatment is shown in table 3, with the standard deviations given in parentheses.

[Insert figures 1-2 and table 3 about here]

Average contributions are highest in the No-Revenge treatments (16.17 per individual from a maximum possible of 20), followed in turn by the Baseline (15.49), the Full Information (10.59) and the Revenge Only (7.21) treatments, but there is considerable heterogeneity between groups in all of the treatments. The treatment differences are consistent with the hypothesized ordering stated in (1). In the No-Revenge Treatment, 5 of 6 groups contribute more than 85% of their endowment over the 20 periods, while no group in the Full Information or Revenge Only treatments does so. On average, contributions in Revenge Only are less than half of the levels in the Baseline and the No Revenge treatments. Figure 1 suggests that after an initial increase in the Baseline and No-Revenge treatments, the average contribution level does not change appreciably as the game is repeated in any treatment, with the exception of Revenge Only, in which it declines.

A Mann-Whitney pairwise statistical test comparing contributions between treatments, maintaining the conservative assumption that each group's activity over the session is a unit of observation, yields the results shown in table 4. The unit of observation is the average contribution of the group over the session (yielding six observations per treatment, one per group), and the null hypotheses are that the median group contributes an identical amount over a twenty-period session. Identical inferences relative to the critical values for $p = .05$ are obtained if the data from the last five periods rather than those from entire sessions are compared.

[Table 4: About Here]

Introducing the possibility of counterpunishment has the effect of reducing contribution levels, regardless of whether or not sanction enforcement is possible. The statement is based on the fact that the difference in contributions between the Baseline and the Revenge Only treatments, as well as the difference between the No-Revenge and the Full Information treatments, is significant. Thus we observe a

[9] Note that, because punishment reduces earnings of the group, group earnings are not necessarily proportional to the

similar effect as Nikiforakis (2004), in that countersanctioning reduces contributions, and find that it generalizes to a setting in which sanction enforcement exists.

Sanction enforcement has a positive, but not significant, effect in increasing contributions. The differences between the Full Information and the Revenge Only Treatments and between the Baseline and the No-Revenge treatments are not significant. The Full Information and the Revenge Only treatments differ only in the ability to sanction based on punishment behavior against others in stage two. The Baseline and the No-Revenge treatments differ from each other in the same manner only. The (borderline, $p < .1$) significant difference between the Baseline and the Full Information treatment indicates that the effect of sanction enforcement is not strong enough to fully offset the decrease in contributions from countersanctioning. Thus, allowing unrestricted reprisals for one round of sanctions reduces overall contributions, though the effect is only of borderline significance.

The ability to engage in sanction enforcement, in conjunction with a prohibition on countersanctioning, is welfare-improving. The difference in total earnings between the Revenge Only and the No-Revenge treatments is significant at the $p < .05$ level, according to a Mann-Whitney rank sum test, with the No-Revenge treatment leading to the higher earnings. The same results are obtained when the last five periods are used in the test rather than all 20 periods. Average earnings per period in the Revenge Only treatment, 19.94 tokens, are comparable to those that would result if no individual made any contributions and there were no sanctions possible (20 tokens). The ability to enforce sanctions on its own does not increase welfare. The No-Revenge treatment does not generate higher earnings than the Baseline treatment. On the other hand, there is some evidence that the possibility of counterpunishment is welfare-reducing. Earnings under the Revenge Only treatment are lower than under the Baseline treatment and the effect is borderline significant. However, a similar test confirms that average group payoffs are not significantly different in the Baseline and Full Information treatments.

Of the four treatments, the greatest number of sanctions are applied in the Baseline treatment, despite the fact that the other treatments include two opportunities to sanction rather than one. The average quantity of sanctions one individual assigns to another in each stage of the four treatments is shown in table 5. The No-Revenge, Revenge-Only and Full Information treatment have a similar overall total number of sanctions applied. Figure 3 illustrates that the number of punishment points assigned decreases during the early periods and stabilizes over the last ten periods of the sessions. The increase in average group earnings over time shown in figure 2 reflects an increase in contributions as well as a decline in sanctions.

[Insert table 5 and figure 3 about here]

---

average group contribution, as they would be if there were no costs associated with the sanctions.

3.2. Who sanctions whom in stage two?

In stage two, agents have observed the contribution decisions of all other individuals, and can condition their sanctioning behavior on this information. Figure 4 illustrates a pattern, which has also been observed by previous authors (Fehr and Gaechter, 2000; Masclet et al., 2003) in the Baseline treatment. The less an individual contributes compared to the group average, the greater the number of punishment points he receives in the second stage of the game (the first sanctioning stage) in all four treatments. In the figure, each bin indicates a range of the recipient's contributions relative to the average level for each treatment. The numbers above the bars indicate the number of observations in the corresponding category. The vertical axis measures the average number of points the individual receives from each of the three other members of his group.

The data in the figure show that contributing within two units of the average draws an average of less than 0.5 punishment points from each individual. Over 50% of all observations are in this range of contributions in each treatment. Contributing considerably less than the group average draws heavy punishment, which increases with the negative deviation of the recipient's contribution from the mean. In all of the treatments, there is a modest increase in the average number of sanctions an individual received the more that his contribution exceeds the group average.

[Insert figure 4 about here]

Falk et al. (2001) and Masclet et al. (2003) also note a strong pattern relating the contributions of those that disbursed and those that received sanctions. The greater the difference between player *i*'s and player *j*'s contribution, provided that *i*'s contribution is higher than *j*'s, the more points that *i* assigns to *j*, holding constant the different between *j*'s contribution and the average. This pattern, as well as the tendency to punish more those who contribute less than the group average, can be seen in table 6. The table reports the result from the following estimation for the data from stage two for the Full Information, No Revenge, and Revenge Only treatments.

$$p_i^{j2t} = \beta_0 + \beta_1 \max\left\{0, c_i^t - c_j^t\right\} + \beta_2 \max\left\{0, c_j^t - c_i^t\right\}$$
$$+ \beta_3 \max\left\{0, \bar{c}^t - c_j^t\right\} + \beta_4 \max\left\{0, c_j^t - \bar{c}^t\right\} \qquad (6)$$

The dependent variable p_i^{j2t} is the quantity of punishment points that player *i* assigns to player *j* in the second stage of period *t*. The variable $\left(\max\left\{0, c_i^{\,t} - c_j^{\,t}\right\}\right)$ is the negative difference between player *j*'s and *i*'s contribution in period *t*. It takes on a value of zero if *j* contributes more than *i,* and a value equal to the difference in *i*'s and *j*'s contributions if *i* contributes more than *j*. A significantly positive coefficient β_1 on this variable would indicate that *i* punishes *j* more the lower *j*'s contribution relative to *i*. The term $\max\left\{0, c_j^{\,t} - c_i^{\,t}\right\}$ has an analogous interpretation as the positive difference between *j*'s and *i*'s contributions in period *t*. A significantly positive β_2 coefficient indicates that *i* punishes *j* more, the more *j* contributed relative to *i*, conditional on *j* contributing more than *i*.[10] The variables $\max\left\{0, \bar{c}^{\,t} - c_j^{\,t}\right\}$ and $\max\left\{0, c_j^{\,t} - \bar{c}^{\,t}\right\}$ are the differences between *j*'s contribution and the group average, conditional on *j* contributing less or more than the group average. A significantly positive coefficient on β_3 (β_4) indicates that *i* punishes *j* more, the farther below (above) the average is *j*'s contribution. In the estimation, each individual pair of players in the same group in each period is the unit of observation.

As shown in table 6, the coefficient β_3 is positive and significant in all three treatments verifying the pattern illustrated in figure 4. Agents receive more punishment, the less they have contributed relative to the group average. The coefficient β_1 is also positive in all three treatments and significant in the Full Information and No Revenge treatments. This indicates that *i* punishes *j* more, the more *i* has contributed relative to *j* in the current period. High contributors are more likely than others to punish low contributors, and the severity of punishment is increasing in the difference between the contribution of the relatively high and low contributors.

[Insert Table 6 About Here]

3.3. Who sanctions whom in stage three?

In the third stage, there are several potential motivations for sanctioning. Agents may wait until the third stage to sanction low contributors,[11] they may enforce sanctions that others failed to apply in stage two, or they may counterpunish. We consider the influences of each of these effects in an estimation of the following equations. Table 7 contains the estimates from the following regression model for the Full Information and the No Revenge treatments:

---

[10] See Cinyabuguma et al. (2004) for a detailed discussion of the punishment of high contributors, which the authors term "perverse" punishment.

[11] The convexity of the cost function for punishment each period means that there are cost savings from spreading out punishment allocations over the two periods. This property, in principle, might encourage greater punishment.

$$p_i^{j3t} = \beta_0 + \beta_1 p_j^{i2t} + \beta_2 \max\left\{0, \sum_{k\neq i} p_j^{k2t} - (\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t})/2\right\} + \beta_3 \max\left\{0, (\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t})/2 - \sum_{k\neq i} p_j^{k2t}\right\}$$
$$+ \beta_4 \max\left\{0, \bar{c}^t - c_j^t\right\} + \beta_5 \max\left\{c_j^t - \bar{c}^t, 0\right\} \quad (7a)$$

For the Revenge Only treatment, the table contains estimates of the equation:

$$p_i^{j3t} = \beta_0 + \beta_1 p_j^{i2t} + \beta_4 \left(\max\left\{0, \bar{c}^t - c_j^t\right\}\right) + \beta_5 \left(\max\left\{0, c_j^t - \bar{c}^t\right\}\right) \quad (7b)$$

The dependent variable in equations (7a) and (7b) is the number of punishment points that player *i* assigns to player *j* in the third stage of period *t*. The coefficient β_1 takes on a positive value if counterpunishment occurs. The coefficient is positive if player *i* reciprocates sanctions he receives by assigning more punishment points to *j* in stage three, the more points *j* assigned to *i* in stage two of the same period. The variable $\sum_{k\neq i} p_j^{k2t} - (\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t})/2$ is the difference between the total number of punishment points that *j* assigned to individuals other than *i* and the average number of punishment points assigned to individuals other than *i* and *j* in stage two. The coefficient β_2 is positive if *i* sanctions *j* more, the more punishment that *j* has disbursed to players other than *i*, relative to the average punishment. If β_3 is positive, sanction enforcement is occurring, since it means that the fewer points *j* assigns relative to the average punishment of third parties in stage two, the more *i* sanctions *j* in stage three. The coefficients β_4 and β_5 capture the dependence of sanctioning behavior in stage three on contribution decisions in stage one, and if punishment of low contributors occurs in stage three, β_4 would take on a positive value.[12] The variables indicating deviations from average punishment of third parties are not included in the Revenge Only treatment, because subjects cannot calculate the number of punishment points *j* has assigned and the relevant average from the information available.

The estimates, reported in table 7, show that counterpunishment, sanction enforcement, and stage three punishment of low contributors all occur. The coefficients β_1 on the variable indicating the number of sanctions assigned in the second stage are positive and significant in all three treatments, indicating the existence of counterpunishment, applied with increasing severity as the initial sanction is increased. This

---

Previous studies indicate that agents punish more, the lower the price of punishment (Anderson and Putterman, 2005; Carpenter, 2005).

[12] An alternative method of capturing the punishment of low contributors is to replace the variables $\max\left\{0, \bar{c}^t - c_j^t\right\}$ and $\max\left\{0, c_j^t - \bar{c}^t\right\}$ with $\max\left\{0, c_i^t - c_j^t\right\}$ and $\max\left\{0, c_j^t - c_i^t\right\}$. Very similar results are obtained if this is done. The coefficients β_1, β_3, and β_4 are all positive and significant at the 1% level in all three treatments. Furthermore, β_0, β_2, and β_5 have identical sign as in table 7 in all three equations.

pattern is also consistent with a pattern of blind vengeance in the No Revenge treatment. Even in the No Revenge treatment, when individuals are not aware of who has sanctioned them, they apparently use information about the sanctions others receive in order to try to avenge the sanctions that they have received in stage two of the game.

[Insert table 7 about here]

The table also shows that sanction enforcement occurs. Players receive more punishment in stage three, the fewer sanctions they assign in stage two compared to the average punishment, once the effect of counterpunishment is taken into account. The evidence is the significantly positive coefficients on β_3 in each of the two treatments. Thus, failure to punish low contributors in stage two at the level others view as appropriate draws punishment in stage three. The positive coefficients on β_4 in all treatments, which are significant in the Full Information and the No Revenge treatments, indicate that low contributions in stage one were also punished in stage three, as in stage two.

Figure 5 illustrates the existence of sanction enforcement. The data are shown for the Full Information and the No-Revenge treatments, in which players have the opportunity to discern who has punished free riders in stage two. The categories in the horizontal axis indicate how much the recipient of sanctions in stage three deviated from average sanctioning behavior in stage two. The vertical axis indicates the average number of punishment points individuals in the different categories received from each other individual. In the Full Information treatment, those who receive the fewest sanctions in stage three are those who sanction close to the group average in stage two. When his deviation from average stage two sanctioning behavior is zero (which occurs only in periods in which all players assign zero sanctions), a player receives on average about 0.2 punishment points in stage 3. However, when the deviation is more than one point from the mean, whether above or below, the average number of points received is 0.6 – 0.8. In the No-Revenge treatment, high punishers are not sanctioned, reflecting the fact that the sanctioning of high punishers in the Full Information treatment consists in part of the exercise of counterpunishment, which is impossible in the No Revenge treatment.

[Insert figure 5 about here]

3.4 Do Stage Two Sanctions Increase Contributions?

Sanctions assigned in stage two have the effect that those applying them presumably intend, to increase subsequent average contributions. On average, the greater the number of sanctions received in a given period, the greater the net increase in contribution to the public good in the subsequent period. Figure 6

illustrates the positive relationship between the total number of sanctions an individual receives in stage two of period t and the average change in his contributions between period t and the next period t+1. The categories in the figure represent different ranges of total number of points received from all group members in the current period. On average contributions decrease if no sanctions are received, and increase if a positive number of punishment points are received. The increase in an individual's contribution is greater the more severe the sanctions he has received in stage two.

[Insert figure 6 about here]

The nature of the relationship between points received in each of the two stages of period t and subsequent contributions in period t+1 is more precisely described in tables 8a and 8b, in which the results of the following estimation are presented.

$$c_i^{t+1} - c_i^t = \beta_0 + \beta_1 \left(\sum_k p_k^{i2t} \right) + \beta_2 \left(\sum_k p_k^{i3t} \right) + \beta_3 \left(c_i^t - \overline{c^t} \right) \qquad (8)$$

In this equation, the dependent variable is the change in contribution between period t and t+1. A positive value of the dependent variable indicates that contributions increase from one period to the next. The coefficient β_1 captures the effect of the number of punishment points received in stage two of period t on the change in contributions. If β_1 is positive, it indicates that the receipt of a larger quantity of sanctions has the effect of inducing a greater subsequent net increase in contributions. While stage two sanctions are presumably unambiguously interpreted as punishment for contribution decisions, stage three sanctions, as we have seen, may reflect other motivations. The extent to which stage three sanctions change subsequent contributions is captured with the coefficient β_2.The deviation from the average contribution, whose effect is captured with β_3, is included as an explanatory variable to account for any regression to the mean that is independent of the number of sanctions received. Such regression to the mean may have any of a number of causes: a consequence of randomness in decisions, a desire to conform to the average, etc.... Table 8a shows the estimates for *low contributors*, those who contribute less than the group average in period t, while table 8b gives the same data for *high contributors*, who contribute more than the group average in period t. The data are separated into high and low contributors because previous work suggests that these two groups may react differently to the receipt of sanctions (Masclet et al., 2003).

[Insert tables 8a and 8b about here]

The estimates show that low contributors who receive more punishment points stage two of period *t*, respond with a more positive net change in contributions for period *t*+1. The coefficient β_1 is significantly positive at the 1% level in all three treatments. Punishment has the intended effect of inducing low contributors to increase their contributions in the next period. However, the same is not the case for high contributors, for which none of the β_1 coefficients is significant at the 5% level. The β_2 coefficients show no general pattern for either high or low contributors, suggesting that receiving sanctions in stage three was not interpreted as punishment for low contributions. For both high and low contributors, the β_3 coefficient is significantly negative in all three treatments, revealing the existence of a general tendency of regression to the mean in contribution levels. The higher one's contribution relative to the average, whether above or below, the stronger the tendency is to lower it in the following period.

3.5. Do counterpunishment and sanction enforcement affect subsequent sanctioning behavior?

In the previous subsection we documented a tendency to increase one's contribution in the next period as a response to the receipt of sanctions in stage two. In this section, we argue that sanction enforcement leads individuals to increase the quantity of sanctions that the recipient assigns in stage two of the following period, while counterpunishment reduces the quantity assigned. Individuals assign more points in stage two of the next period, the more sanctions they receive in the third stage of the current period in the Full Information and the No Revenge treatments. In those treatments, the receipt of points can be interpreted as sanction enforcement. However, the opposite is true in the Revenge Only treatment, in which sanctions received in stage three may only be interpreted as counterpunishment. The more aggressively the sanctions one assigns in a period are avenged, the fewer sanctions one assigns in the next period.

Figure 7 illustrates these effects for *low punishers*, those who sanction less than the average in their group in stage two of period *t*. Each bin represents a range of sanctions an individual receives in the third stage of period *t*. The vertical axis is the net change from period *t* to *t*+1 in the total number of punishment points an individual assigns in stage two. This change is increasing in the number of points received for the Full Information and the No Revenge treatments, but not in the Revenge Only treatment. The same effect appears in tables 9a and 9b, which display the results of the estimation of equation (9), for low and high punishers, respectively.

$$\sum_k p_i^{k,2,t+1} - \sum_k p_i^{k,2,t} = \beta_0 + \beta_1 \sum_k p_k^{i3t} + \beta_2 \left(\sum_k p_k^{i2t} - \overline{\sum_k p_k^{2t}}\right) \qquad (9)$$

The dependent variable in the equation is the change in the total amount of punishment that player *i* assigns between stage 2 of periods *t* and stage 2 in period *t*+1. The independent variables are the sum of

points the individual has received in the stage three of period t and the difference between the number of points he assigns and the average number of points individual members of the group assign in period t. If $\beta_1 > 0$, agents respond to sanction enforcement or to counterpunishment with increases in the quantities of sanctions they assign in stage 2 of the following period. If $\beta_2 < 0$, there if a tendency for those who have sanctioned less relative to the average in stage two of period t, to exhibit a greater net increase in the sanctions they assign in stage two of period t+1 relative to stage two of period t. We divide the data to distinguish between low and high punishers. A low (high) punisher in period t is an individual who distributed fewer (more) punishment points in stage two of period t than the average in her group. While the mean sanction assigned in stage two is only known to individuals in the Full Information treatment, agents can compare the total quantity of sanctions they assign to the total number they receive and have a rough indication of how they compare to the average.

The estimates show that in the Full Information and No-Revenge treatments, the greater the number of sanctions a low punisher receives in stage three of period t, the greater the net increase in the number of punishment points he distributes in stage two of period t+1 relative to period t. He acts as if he has interpreted the punishment he has received as sanction enforcement, and responds as if to reduce the receipt of future sanction enforcement. No such effect is observed for high punishers, who do not appear to interpret stage three sanctions they receive as punishment for insufficient assignment of sanctions. The reverse is true under the Revenge Only treatment, in which low punishers cannot be identified, and stage three sanctions are interpreted as counterpunishment. Under Revenge Only, the more sanctions that one receives in stage three of period t, the fewer one assigns during stage two of period t+1. This effect is observed for both low and high punishers. Thus, the use of counterpunishment in stage three for prior sanctions has the effect of deterring the sanctioner in the next period.

[Insert figure 8 and tables 9a and 9b about here]

After the above effects are taken into account, the more that individuals punished in excess of the average sanction in the second stage in a given period, the greater the tendency to sanction less in the following period. This effect is observed in all three treatments for high sanctioners, as can be seen from the negative and significant β_2 coefficients in each of the three treatments for high punishers. No similar overall tendency toward (or away from) conformity is detected for low punishers.

3.6. The Six Stage Full Information treatment

The decrease in contributions that the introduction of a second stage of punishment induces suggests that additional stages of unrestricted punishment might further reduce contributions. This question can be

considered with a comparison of behavior in the Six-Stage Full Information with the Baseline treatments, which differ from each other only in the existence of additional stages of punishment.

Contributions are indeed lower in the 6SFI treatment than in the Baseline treatment. Figure 1 indicates that average contributions are lower throughout the time horizon of the sessions in 6SFI than in the Baseline treatment. The results of a Mann-Whitney rank sum test for treatment differences, given in Table 4, reveal that the differences in median group contributions between 6SFI and Baseline are significant at the 5% level.

The effect on welfare of the additional stages of punishment opportunities is large and negative. Figure 3 illustrates the effect. In the early periods of the 6SFI treatment, welfare is negative on average. In later periods it increases, but remains below the level in all other treatments, except for the Revenge Only treatment. Indeed, average earnings for the group remain below 80 over the entire session, indicating that earnings are less than at a benchmark where contributions are zero from all players and no punishment is possible. As table 4 indicates, the welfare level over the entire 20 periods is significantly lower in 6SFI than in each of the other four treatments.

The source of the lower welfare is twofold. While contributions are lower in the 6SFI than in the Baseline and the Full Information treatments, the number of sanctions applied is also higher in 6SFI. Table 4 displays the data, indicating that the average number of sanctions an individual receives in a period is equal to 4.11 points. This represents a reduction of 41.1% of earnings after the first stage, not including the costs the sanctioners incur. The number of points assigned is 2.71 times the amount in the next highest treatment. The number of points assigned within a period follows a distinct pattern of small declines in stages 2 – 5, and a large increase in stage 6. Some of this activity in the last punishment stage appears to consist of sanctions for earlier contribution or punishment decisions that have been delayed until they are immune from counterpunishment.

In some periods, a phenomenon of escalating counterpunishment is observed. This phenomenon consists of a sanction that player *i* applies to *j*, followed by the assignment of counterpunishment by *j* to *i*, and one of more reciprocal reprisals. We give two example of this phenomenon here. In period 3, players A, B, C, and D in group 1 contribute 12, 8, 12, and 0 tokens, respectively. Player A then assigns 5 points to D in stage two. D responds by assigning 1 point to A in stage three. Then A assigns 3 to D, D assigns 2 to A, and A assigns 2 to D in the next three stages. Similarly, in period 6, the contributions of players A – D are 12, 10, 5, and 15 tokens respectively. In stage two, D allocates one point to C and C responds by assigning D one point in stage three. D assigns C two points, C gives one point to D, and D directs two points to C, respectively, in rounds 4 – 6.

4. Conclusion

In this study, we investigate the impact of allowing punishment of sanctioning behavior on decisions and payoffs of individuals facing a social dilemma. Our experimental design closely follows the structure of Fehr and Gaechter (2000). The results of our study show that the existence of additional rounds of sanctions, in which any player may sanction any other, has a significantly negative effect on the level of contributions. Our Six Stage Full Information treatment yields significantly lower contribution levels than a system with only one round of sanctions. The lower contributions, in conjunction with higher sanctioning levels, leads to welfare levels below those when only one round of sanctions exists. Indeed, if five rounds of sanctions exist, welfare is lower than when no sanctioning mechanism exists and contributions are zero for each individual. Thus, the availability of four additional rounds of sanctions more than completely negates any improvement in welfare that a single sanctioning opportunity generates. These results underscore the idea that anonymity of sanctioning from reprisals is crucial for the operation of a successful sanctioning mechanism. Voluntary contributions are highly susceptible to the free-rider problem, while the anonymous sanctioning environment of our Baseline treatment or of Fehr and Gaechter (2000) are highly conducive to cooperation. Setting with repeated opportunities to punish appear to generate contribution levels that lie in between these two extreme cases.

Our treatments with two rounds of sanctions allow us to study the effect of opportunities to engage in counterpunishment and in sanction enforcement in isolation. When counterpunishment is possible, sanctioning is deterred. As a result, the deterrent effect on free-riding of the threat of sanctions is mitigated and contribution levels fall. The bases for this statement are the fact that the Revenge Only treatment exhibits lower contribution levels that the Baseline treatment, and the fact that the Full Information treatment generates lower contributions than the No Revenge treatment. The only possible explanation for these effects is that the opportunity to countersanction reduces contributions. Our results are consistent with those reported in the recent work of Nikiforakis (2004), who finds that when agents are permitted to counterpunish, but not to enforce sanctions, initial sanctioning is deterred and contributions to the public good decrease as a result.

However, when sanction enforcement is possible, contributions change in the opposite direction, although the effect is not significant. In the Full Information treatment, agents are able to punish those failed to sanction free riders as well as to exact revenge on those who have sanctioned them. The effect is to increase average contributions by 50% over the level in the Revenge Only treatment, partially offsetting the reduction in contributions due to counterpunishment. There is no tendency for contributions to decline with repetition of the game in any treatment where sanction enforcement is possible (namely the Full

Information, No Revenge and Six Stage Full Information treatments). However, in the Revenge Only treatment, in which sanction enforcement is not possible, average contributions decline over time.

The sanctions operate in an intuitive manner at the individual level. Agents sanction low contributors in the second stage. In the third stage they sanction low contributors and low sanctioners, as well as counterpunish. Sanctions received in the second stage increase recipients' contributions in the following period. In the third stage, counterpunishment reduces the quantity of sanctions recipients assign in the following period, while sanction enforcement increases it. The overall effect of a second stage of punishment and full observability of prior contribution and punishment decisions is a (borderline significant) reduction in contributions, as the effect of counterpunishment is more powerful than the effect of sanction enforcement. Additional rounds of sanctioning opportunities appear to further erode contribution levels. The 6SFI treatment shows average contributions significantly lower than those of the Baseline treatment.

The experiment provides further evidence of the formation of social norms on the part of a group. The agents who are sanctioned are not only free riders in the contribution phase, but those who deviate from implicit group norms of punishment in a direction viewed as opportunistic. The addition of multiple rounds of sanctions illustrates the establishment of a norm of punishment along with a norm of contribution and demonstrates that at least some agents are prepared to pay from their earnings to enforce the norm.

7. Bibliography

Anderson C. and L. Putterman. 2005. "Do Non-Strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contributions Mechanism". *Games and Economic Behavior*, forthcoming.

Andreoni, J. 1988. "Why Free Ride: Strategies and Learning in Public Goods Experiments", *Journal of Public Economics*, 35 (1), pp. 57-73.

Bochet, O., T. Page, and L. Putterman. 2005. "Communication and Punishment in Voluntary Contribution Experiments", *Journal of Economic Behavior and Organization*, forthcoming.

Carpenter, J. 2005. "Punishing Free Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods*." Games and Economic Behavior*, forthcoming.

Carpenter, J. 2005. "The Demand for Punishment." *Journal of Economic Behavior and Organization,* forthcoming.

Cinyabuguma M., T. Page and L. Putterman. 2004. "On Perverse and Second-Order Punishment in Public Goods Experiments with Decentralized Sanctioning", working paper, Brown University.

Falk, A., Fehr, E. and U. Fischbacher. "Driving Forces of Informal Sanctions", *Econometrica*, forthcoming.

Fehr E. and S. Gaechter 2000. "Cooperation and Punishment in Public Goods Experiments". *American Economic Review*, September, 90, 4, pp. 980-94.

Fischbacher U. 1999. "z-Tree: A Toolbox for Readymade Economic Experiments," working paper, University of Zurich, Institute for Empirical Research in Economics.

Gaechter S. and E. Fehr. 1999. "Collective Action as a Social Exchange." *Journal of Economic Behavior and Organization*, 39 (2), pp. 341-69.

Gaechter S. and B. Herrmann. 2005. "Norms of Cooperation Among Urban and Rural Dwellers: Experimental Evidence from Russia", mimeo, Harvard University and the University of Nottingham.

Isaac, R. M., K. McCue, and C. Plott. 1985. "Public Goods Provision in an Experimental Environment", *Journal of Public Economics*, 26 (1), pp. 51 – 74.

Isaac, R. M., and J. Walker. 1988a. "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism," *Quarterly Journal of Economics* 103 (1), pp.179-99.

Isaac, R. M., and J. Walker. 1988b. "Communication and Free-Riding Behavior: The Voluntary Contributions Mechanism," *Economic Inquiry* 26(4), pp. 585-608.

Kiyonari, T., and P. Barclay. "Second Order Punishment and Reward in Public Goods Games", mimeo, McMaster University.

Ledyard J. 1995. “Public Goods: A survey of experimental research”, in Kagel J. and Roth. A., eds., Handbook of experimental economics. Princeton, Princeton University Press, pp. 111-94.

Masclet, D., C. Noussair, S. Tucker and M. Villeval. “Monetary and Non-Monetary Punishment in the Voluntary Contributions Mechanism.” *American Economic Review*, 93 (1), pp. 366-380.

Nikiforakis N.S. 2004. "Punishment and Counter-punishment in Public Goods Games: Can We Still Govern Ourselves?" working paper, Royal Holloway University of London.

Ostrom, E., J. Walker and Gardner, R. 1992. “Covenants with and without a Sword: Self-governance is possible”. *American Political Science Review*, 86 (2), pp. 404-17.

Yamagishi, T. 1986. The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51(1) pp. 110-16.

Table 1: Treatments and Anticipated Effects on Contributions

Treatment	Punishment for Low Contributions	Counter-Punishment	Sanction Enforcement
B	+	N/A	N/A
RO	+	-	N/A
NR	+	N/A	+
FI	+	-	+

Table 2 – Punishment levels and associated costs for the sanctioner

$\#\,points$	0	1	2	3	4	5	6	7	8	9	10
Cost $k_i(p_i^j)$	0	1	2	4	6	9	12	16	20	25	30

Table 3: Average Individual Contribution Levels by Group in each Treatment

	Baseline	Full Information	Revenge Only	No Revenge	6 Stage Full Information
Group 1	11.5	12.93	6.225	7.47	12.33
	(3.44)	(4.20)	(2.42)	(4.38)	(5.41)
Group 2	16.73	3.42	16.01	18.85	4.48
	(4.71)	(5.75)	(6.15)	(3.13)	(5.35)
Group 3	17.06	12.65	2.275	18.85	16.05
	(4.79)	(3.46)	(2.59)	(3.24)	(5.13)
Group 4	18.03	5.575	2.15	17.31	10.4
	(2.504)	(6.37)	(1.51)	(5.39)	(3.89)
Group 5	10.63	12.52	5.45	17.33	5.49
	(3.40)	(3.52)	(3.20)	(4.87)	(2.09)
Group 6	18.96	16.47	11.13	17.23	10.89
	(2.14)	(5.13)	(6.68)	(4.03)	(3.00)
Average	15.485	10.594	7.206	16.17	9.93
Std. Dev.	(3.497)	(4.73833)	(3.75)	(4.173)	(5.87)

Table 4: Results of Mann-Whitney Rank Sum Tests of Differences in Contribution Levels and Earnings Between Treatments

(Level of confidence at which null hypothesis of no differences between treatments can be rejected, each session mean is a unit of observation)

Contributions				
	Full information	**Revenge Only**	**No revenge**	**6 stages Full Information**
Baseline	***p* < .10**	***p* < .01**	**Not Sig.**	***p* < .05**
Full Inf	**___**	**Not Sig.**	***p* < .01**	**Not sig.**
Revenge	**___**	**___**	***p* < .005**	**Not sig.**
No revenge	**-----**	**-----**	**------**	***p* < .05**
Earnings				
Baseline	**Not Sig.**	***p* < .1**	**Not Sig.**	***p* < .02**
Full Inf	**___**	**Not Sig.**	**Not Sig.**	***p* < .05**
Revenge	**___**	**___**	***p* < .05**	***p* < .05**
No revenge	**-----**	**-----**	**------**	***p* < .01**

Table 5: Average Quantity of Sanctions ($\sum_i \sum_k p_i^{kmt}$ /n) Assigned by Individuals to the Rest of the Group in a Period: All Treatments

Treatment	Baseline	Full Information	No Revenge	Revenge Only	Six Stage Full Information
Average points assigned in stage 5					0.769
Average points assigned in stage 4	/	/	/	/	0.650
Average points assigned in stage 3	/	/	/	/	0.544
Average points assigned in stage 2	/	0.57	0.37	0.38	0.527
Average points assigned in stage 1	1.512	0.46	0.65	0.73	1.617
Average assigned over the stages	1.512	1.03	1.02	1.11	4.11

Table 6 : Sanctions Assigned by *i* to *j* in Second Stage as a Function of Contribution Decisions of *j* in Stage One of Current Period

$$p_i^{j2t} = \beta_0 + \beta_1 \max\left\{0, c_i^t - c_j^t\right\} + \beta_2 \max\left\{0, c_j^t - c_i^t\right\} + \beta_3 \max\left\{0, \bar{c}^t - c_j^t\right\} + \beta_4 \max\left\{0, c_j^t - \bar{c}^t\right\}$$

	Full Information	Revenge Only	No Revenge
Constant	-5.02***	-3.59***	4.68***
	(0.308)	(0.199)	(0.26)
Amount sanctioner contributed above recipient's contribution (β_1)	0.1943***	0.059	0.181***
	(0.039)	(0.038)	(0.04)
Amount sanctioner contributed below recipient's contribution (β_2)	-0.147*	0.289***	0.068
	(0.064)	(0.031)	(0.047)
Amount recipient contributed below average (β_3)	0.29***	0.176***	0.336***
	(0.047)	(0.042)	(0.048)
Amount recipient contributed above average (β_4)	0.1617*	-0.327***	0.022
	(0.070)	(0.053)	(0.061)
Log likelihood	-1144.49	-1723.84	-1346.43
Pseudo R2	0.126	0.059	0.135
Observations	2880	2880	2880

*** 1% significance level, ** 5% significance level, * 10% significance level, Tobit estimation used, Standard errors are in parentheses

Table 7: Number of Punishment Points that Player *i* Assigns to *j* in the Third Stage as a Function of Prior Contribution and Sanctioning Decisions of Recipient

$$p_i^{j3t} = \beta_0 + \beta_1 p_j^{i2t} + \beta_2 \max\left\{0, \sum_{k\neq i} p_j^{k2t} - (\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t})/2\right\} + \beta_3 \max\left\{0, (\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t})/2 - \sum_{k\neq i} p_j^{k2t}\right\}$$

$$+ \beta_4 \max\left\{0, \bar{c^t} - c_j^t\right\} + \beta_5 \max\left\{c_j^t - \bar{c^t}, 0\right\}$$

	Full Information	Revenge Only	No Revenge
Constant	-3.93*** (0.230)	-4.146*** (0.280)	-4.43*** (0.29)
Points *j* assigned to *i* in second stage (β_1) (Counterpunishment)	0.726*** (0.112)	1.209*** (0.095)	0.704*** (0.089)
Positive deviation of recipient from average punishment in second stage (β_2)	0.346*** (0.127)		-0.217 (0.151)
Negative deviation of recipient from average punishment in second stage (β_3) (Sanction Enforcement)	0.712*** (0.157)		0.408*** (0.147)
Amount recipient contributed below the average (β_4) (Punishment of Low Contributors)	0.263*** (0.028)	0.0160 (.0267)	0.316*** (0.0274)
Amount recipient contributed above the average (β_5)	-0.085** (0.039)	-.0380 (.0365)	0.0464 (0.035)
Observations	2880	2880	2880
Log likelihood =	-1469.38	-1047.55	-1014.02
Pseudo R2 =	0.0623	0.09	0.1077

*** 1% significance level, ** 5% significance level, * 10% significance level,

Table 8a: The Effect of Period *t* Sanctions on Changes in Contribution Between Periods *t* and *t*+1: Low Contributors

$$c_i^{t+1} - c_i^t = \beta_0 + \beta_1\left(\sum_k p_k^{i2t}\right) + \beta_2\left(\sum_k p_k^{i3t}\right) + \beta_3\left(c_i^t - \overline{c^t}\right)$$

	Baseline	Full Information	Revenge Only	No Revenge
Constant (β_0)	-0.783***	0.556***	-1.887***	.4633**
	(0.1409)	(0.71)	(.1693)	(.1863)
Points received in second stage of period *t* (β_1)	0.1407***	.6109***	.4353***	.2198***
	(0.043)	(.0734)	(.0921)	(.0943)
Points received in third stage of period *t* (β_2)		-.1461*	.2239**	0892
		(.0832)	(.1168)	(.1324)
Deviation from Average Contribution in period *t* (β_3)	-0. 85***	-.2785***	-.8374***	-0.37***
	(0.031)	(.0365)	(.0326)	(0.05)
R^2	0.51	0.170	0.47	0.182
Observations	1098	1182	1218	828

*** 1% significance level, ** 5% significance level, * 10% significance level,
Standard errors are in parentheses

Table 8b : The Effect of Period *t* Sanctions on Changes in Contribution Between Periods *t* and *t*+1 : High Contributors

$$c_i^{t+1} - c_i^t = \beta_0 + \beta_1\left(\sum_k p_k^{i2t}\right) + \beta_2\left(\sum_k p_k^{i3t}\right) + \beta_3\left(c_i^t - \overline{c^t}\right)$$

	Baseline	Full Information	Revenge Only	No Revenge
Constant (β_0)	-0.2536	.5023***	.4438***	1.0967***
	(0.167)	(.1689)	(.1650)	(.1836)
Points received in second stage of period *t* (β_1)	-0.072203	.1052	-.0725	-.0939
	(.0654)	(.1132)	(0.096)	(.1020)
Points received in third stage of period *t* (β_2)		.1966**	-.7655***	.2115
		(.1015)	(.133)	(.1609)
Deviation from Average Contribution in period *t* (β_3)	-0.373***	-.6758***	-.6219***	-0.459***
	(.0422)	(.0356)	(0.035)	(0.036)
R^2	0.055	0.23	0.209	0.171
Observations	1344	1218	1380	840

*** 1% significance level, ** 5% significance level, * 10% significance level,
Standard errors are in parentheses

Table 9a: The Effect of Stage Three Punishment on Sanctions Assigned in the Second Stage of Following Period: Low Punishers

$$\sum_k p_i^{k,2,t+1} - \sum_k p_i^{k,2,t} = \beta_0 + \beta_1 \sum_k p_k^{i3t} + \beta_2 (\sum_k p_k^{i2t} - \overline{\sum_k p_k^{2t}})$$

	Full Information	Revenge Only	No Revenge
Constant	.3508***	.2245***	.1906***
	(.1300)	(.0545)	(.0614)
Points received in third stage of period t	.1536**	-.1278*	.0796**
	(.0655)	(.069)	(.0308)
Deviation from Average punishment in period t	-.1109	-.0400	.0717
	(.2029)	(.0840)	(.0979)
R^2	0.012	0.005	0.010
Observations	498	720	660

*** 1% significance level, ** 5% significance level, * 10% significance level, Standard errors are in parentheses

Table 9b: The Effect of Stage Three Punishment on Sanctions Assigned in the Second Stage of Following Period: High Punishers

$$\sum_k p_i^{k,2,t+1} - \sum_k p_i^{k,2,t} = \beta_0 + \beta_1 \sum_k p_k^{i3t} + \beta_2 (\sum_k p_k^{i2t} - \overline{\sum_k p_k^{2t}})$$

	Full Information	Revenge Only	No Revenge
Constant	.5445***	.6735***	.3246**
	(.11505)	(.1128)	(.145)
Points received in third stage of period t	.0875	-.4840***	-.0325
	(.0556)	(.0553)	(.0829)
Deviation from Average punishment in period t	-.9997***	-.4510***	-.6184***
	(.0367)	(.0364)	(.0418)
R^2	.5839	. 3415	0.272
Observations	564	690	600

*** 1% significance level, ** 5% significance level, * 10% significance level, Standard errors are in parentheses

Figure 1: Average Individual Contribution Levels in Each Treatment

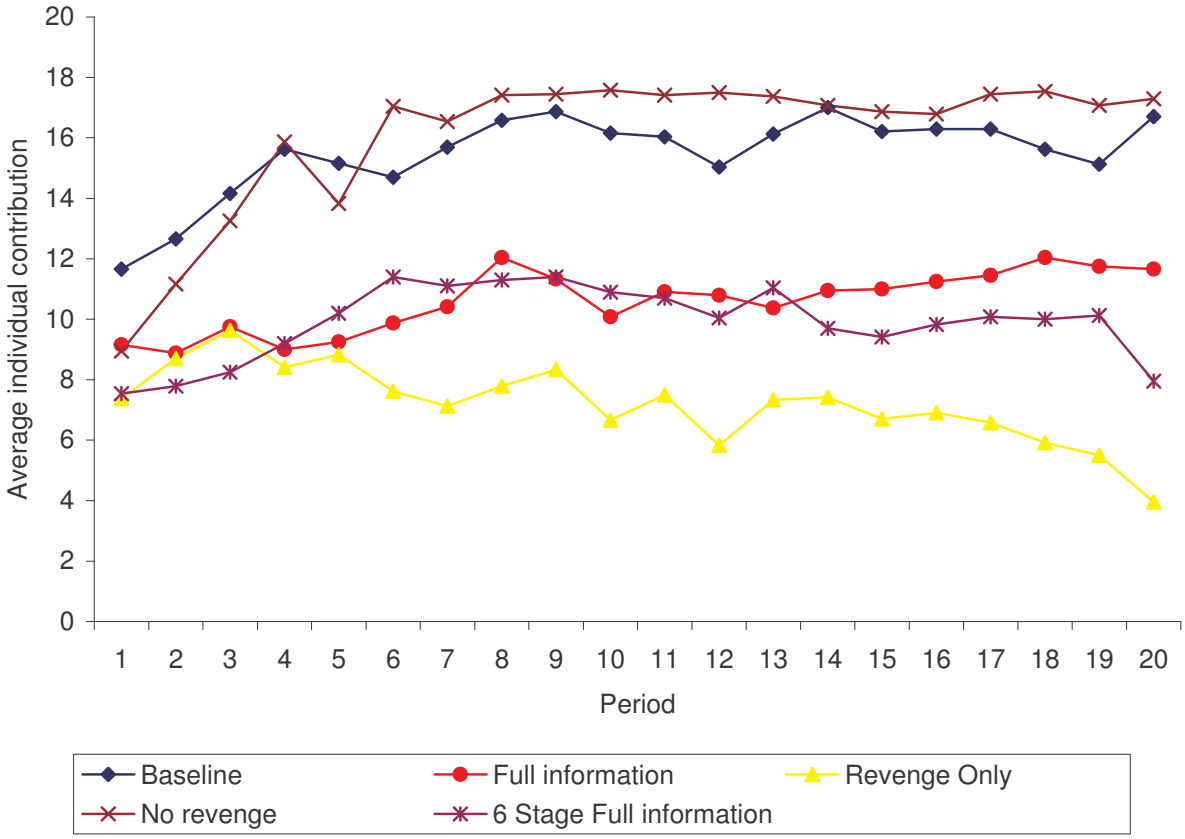


Figure 2: Average earnings per group in each treatment, by period

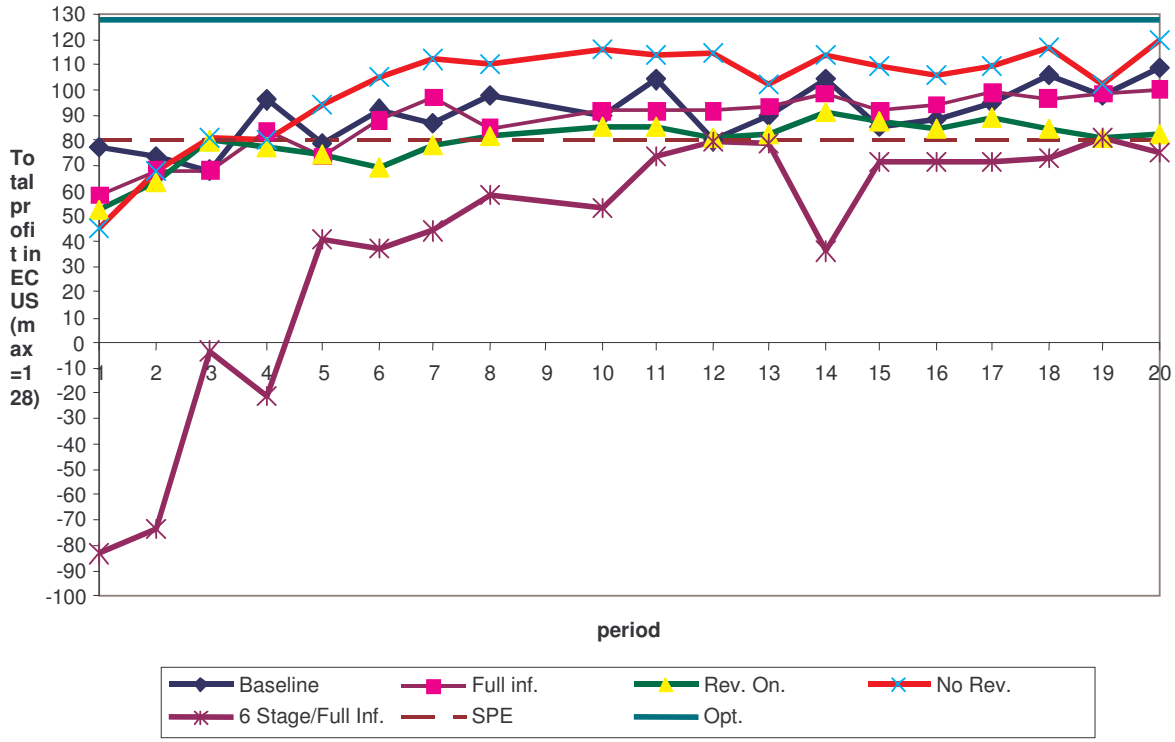


Figure 3 : Average Quantity of Punishment Points Assigned from One Individual to Another, By Period, B, FI, NR, and RO Treatments

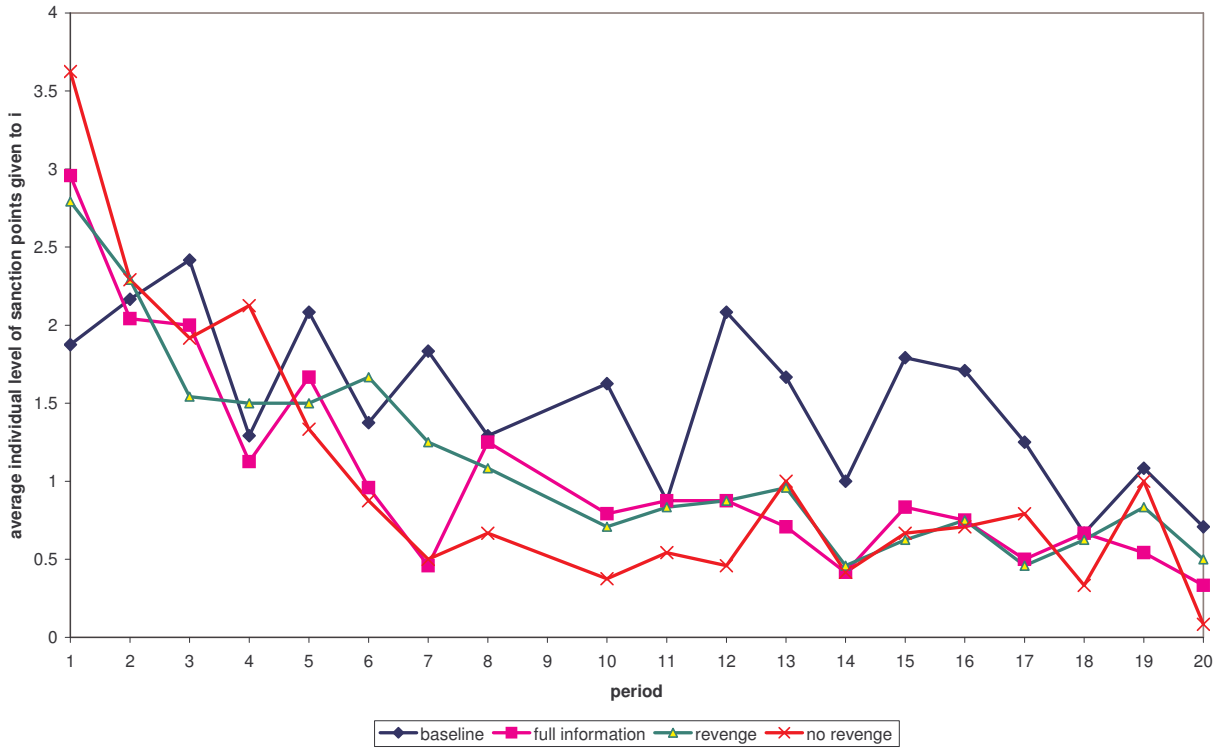


Figure 4: Average Number of Punishment Points Assigned in Stage Two as a Function of Recipient's Stage One Contribution

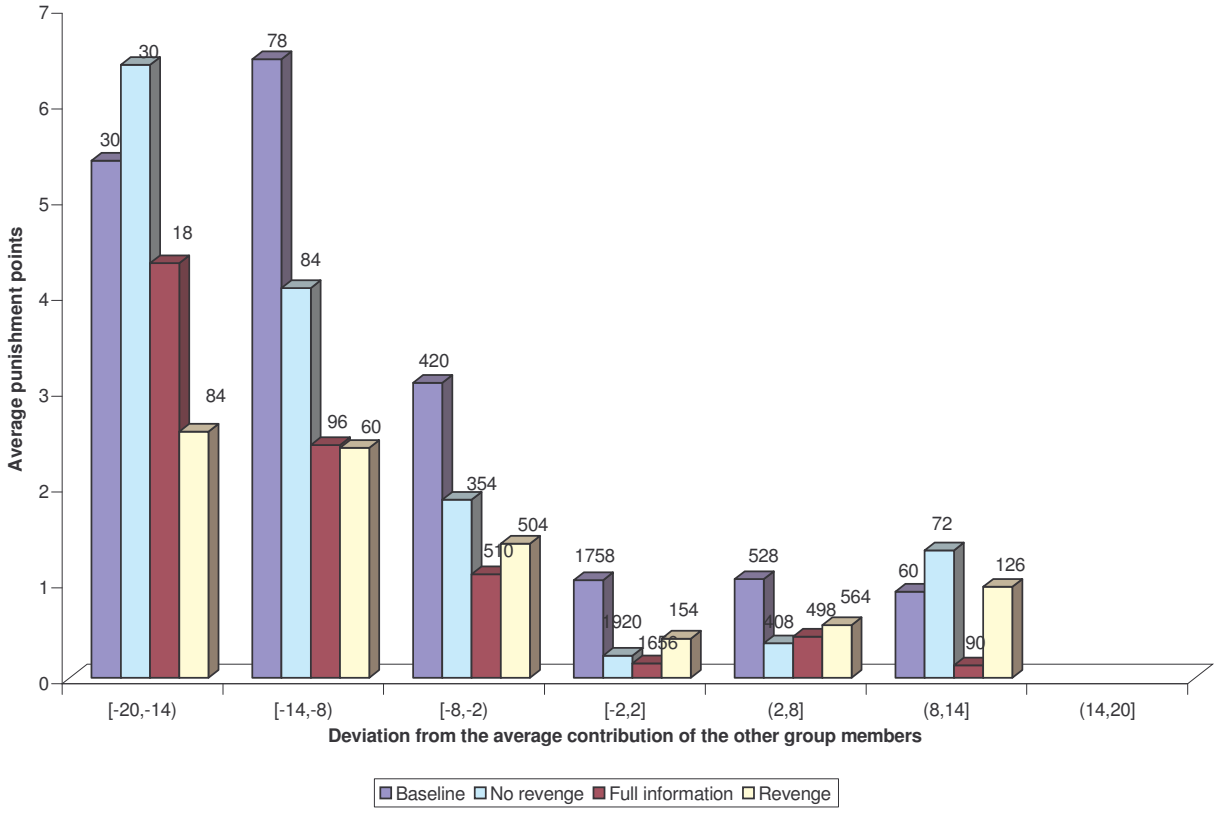


Figure 5. Average Number of Punishment Points Received in Stage Three as a Function of the Difference Between Own and Average Points Assigned in the Second Stage

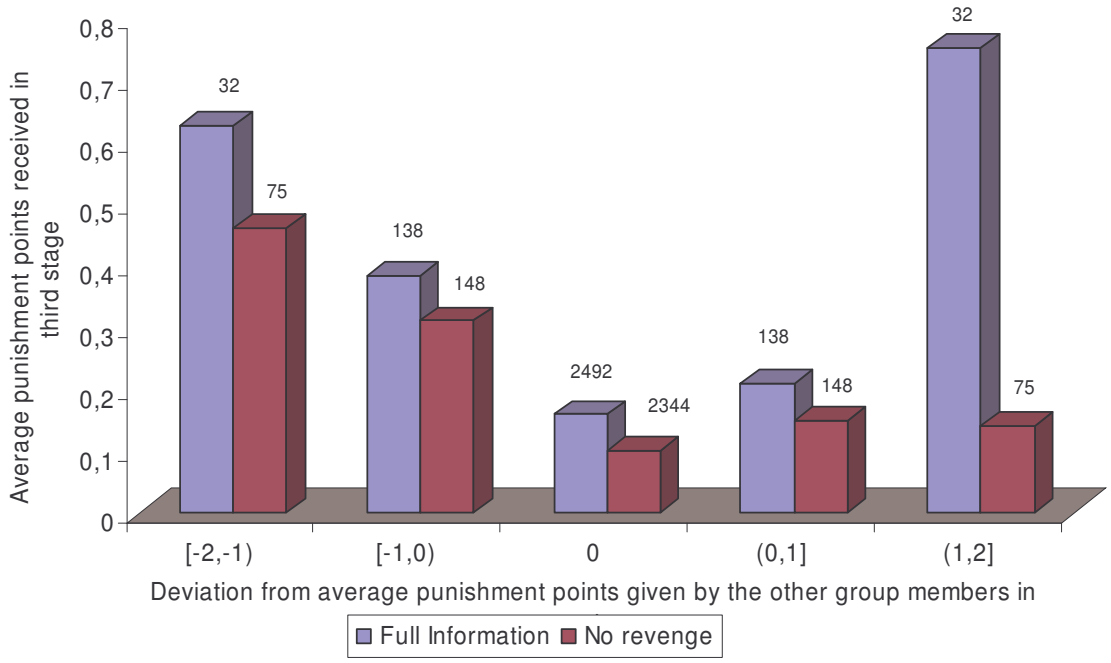


Figure 6. The Effect of the Number of Punishment Points Received in Stage Two of Period *t* on the Change in Contribution Between Periods *t* and *t*+1

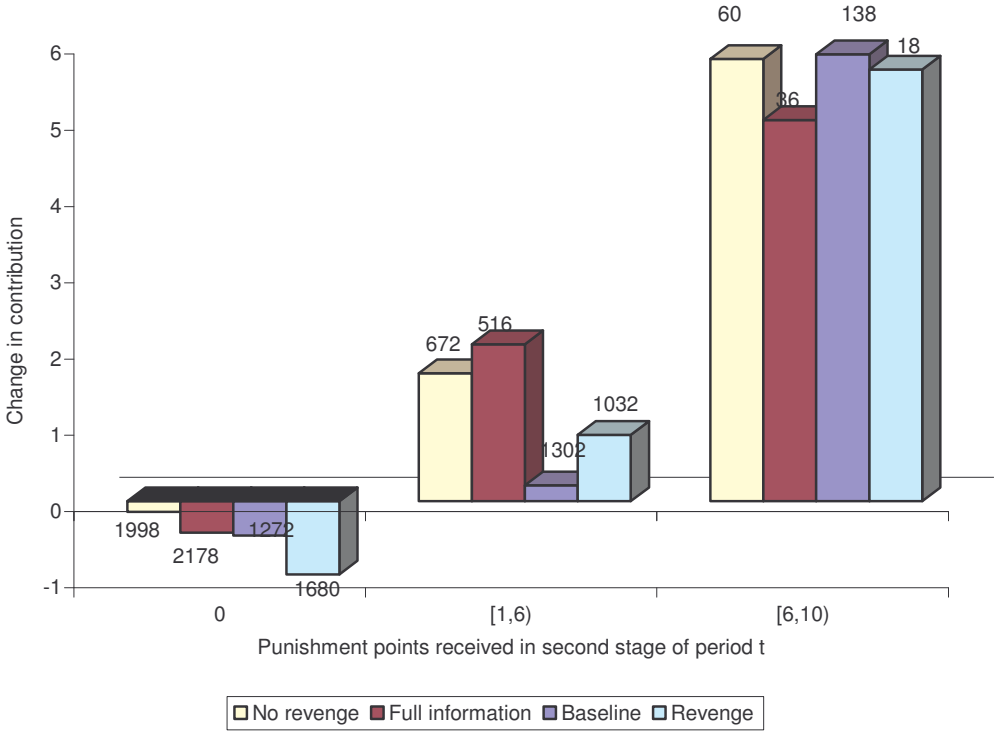
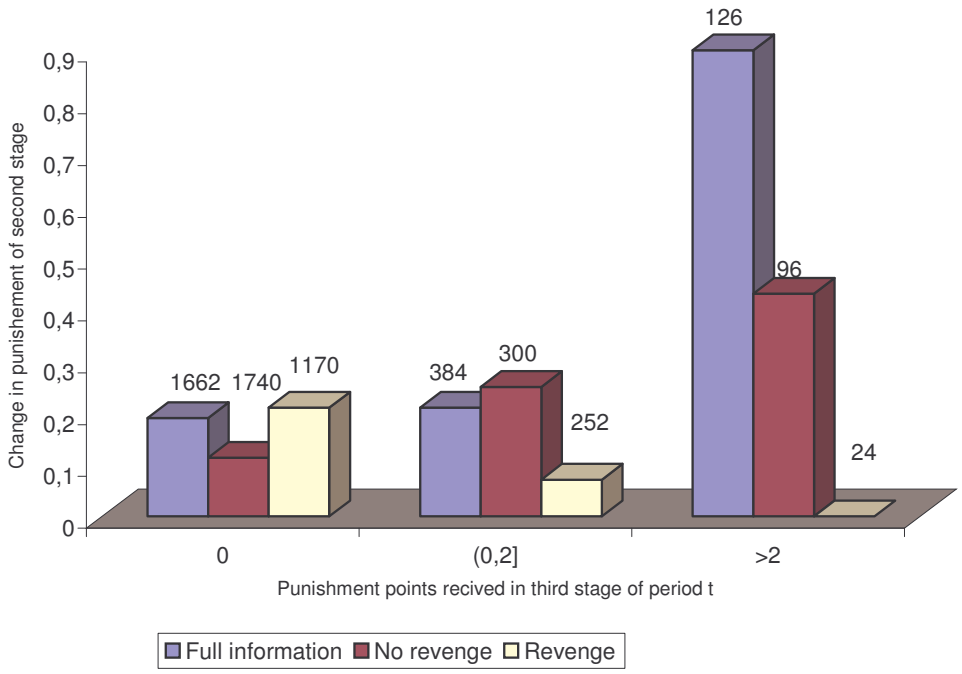


Figure 7 : The Effect of Sanctions Received in the Third Stage of Period *t* on the Change in Punishment Points Assigned between the Second Stages of Periods *t* and *t*+1, Low Punishers



<<<Note: The following pages contain a translation from the original French text of the instructions used for the Full Information treatment. The instructions for the other four treatments involve only minor changes from the instructions included here, reflecting in the differences in the information available to players after stage two, and the number of punishment stages in the game.>>>

GENERAL INSTRUCTIONS

You are participating in an economic experiment, during which you can earn a considerable amount of money. Your earnings depend on your decisions and the decisions of the other participants in the experiment. Therefore, it is important to read these instructions carefully.

The instructions that we have distributed to you are your private information. It is forbidden for you to communicate with the other participants during the experiment. If you do, you will be excluded from the session and you will not receive any monetary payment.

All of the transactions in the experiment and your earnings will be calculated in terms of ECU (Experimental Currency Units). Your earnings in ECU for this experimental session will be converted to Euros and paid to you in cash on the following basis:

-Your final earnings in terms of ECU equal the total earnings from all of the rounds of play that make up the session.

- These final earnings in ECU will be converted to Euros on the following basis: 1 ECU is worth 0,02 Euros.

- In addition, you will receive a participation fee of 8 Euros.

Activity in Each Round

This experimental session is composed of a sequence of 20 rounds. In each round, participants are divided into groups of four. Therefore, you will be grouped with three other people. **For the entire session, you will be grouped with the same participants**. You will not know the identity of the other members of your group. Each round will have three stages.

During the first stage, you must decide how many ECUs that you would like to contribute to a "project". In the second stage, you are informed of the contributions that the three other members of your group made to the project. You may then choose to reduce or not to reduce any other group member's earnings by assigning points to him/her. Finally, in the third stage, you are informed of the number of points that each player assigned to each other player, including who assigned points to you. You then have another opportunity to assign points to any of the group members. The following paragraphs describe the game in detail.

- **The first stage**

  - At the beginning of each round each individual receives 20 ECU.

  - Each of the four members of a group, including you, decide simultaneously on how many of these ECU, 0 to 20 inclusive, that he would like to contribute to the project. As soon as you make your decision, please click on the [OK] button. After you have selected [OK], you may not change your contribution decision for the current round.

- At the end of stage one, you will be informed individually, on your computer screen, of your current earnings after the first stage, which consists of two elements:

- (1) The amount of your initial 20 ECU that you have kept for yourself (that is, 20 ECU – Your Contribution to the Project).
- (2) Your income from the project. The income to you is equal to 40% of the total of the four individual contributions to the project.
- Your earnings at the end of the first stage are calculated by the computer in the following manner:

Your earnings at the end of the first stage = (20 – your contribution to the project) + 40% (total group contribution to the project)

The earnings of each member of your group are calculated in the same manner, which means that each individual receives the same income from the project. Suppose for example, that the total number of ECU the group contributes is 60 ECU. Then, each member of the group receives an income from the project equal to 40% of 60 ECU, which equals 24 ECU. As an another example, if the total of the contributions to the project is 9 ECU, each member of the group receives an income from the project equal to 40% of 9 ECU, which equals 3.6 ECU.

Each ECU from your initial amount of 20 that you keep for yourself gives you 1 ECU in earnings. However, if you decide to assign an ECU to the project instead, the total contribution to the project increases by 1 ECU. Your income therefore increases by 40% of 1 ECU, which equals 0.4 ECU. The income of each other member of your group also increases by 0.4 ECU, so the total income of the group increases by 0.4*4 ECU = 1.6 ECU. This means that increasing your contribution to the project increases the total income to the group.

In the same manner, you receive income from every ECU that any other member of your group assigns to the project. For each ECU contributed by another member of the group, you earn .4 ECUs (40% of 1 ECU).

- **At the beginning of the second stage**, your screen displays the amount that each of the other members of your group contributed to the project in stage 1. Each of you will then have the **possibility of reducing or leaving unchanged the earnings of each of the other members of your group by assigning points. You can assign these points to demonstrate a level of disapproval (10 points for the most severe disapproval, 0 points for the absence of disapproval). Each point assigned to an individual reduces the recipient's earnings by 10% of his stage 1 earnings.**

Similarly, your earnings may be reduced if the other members of your group assign points to you.

First, you are informed of the amount that each of the three other players has contributed to the project in the 1st stage of the game in the current round. Attention: the order in which the decisions of the three other members of your group appear on your screen randomly change from round to round (said differently, the number that appears first on your screen does not always correspond to the decision of the same player).

Afterwards, you must decide on the number of points that you will assign to each of the three other members of your group to reduce their earnings or leave their earnings unchanged. You must enter a number of points for each subject between 0 and 10 inclusive. If you do not wish to reduce an individual's earnings, enter the number 0.

If you assign points, you incur a cost that depends on the number of points distributed to each subject. The greater the number of points you assign, the higher your costs are. Your

total cost is equal to the total of the three individual costs you pay to assign points to each of the other group members. The table below shows the relationship between the number of points you assign to an individual and the associated cost to you.

Number of points you assign	0	1	2	3	4	5	6	7	8	9	10
Cost of points to you	0	1	2	4	6	9	12	16	20	25	30

If you, for example, assign 2 points to one member of your group, it costs you 2 ECU. If you assign 9 points to another group member, it costs you an additional 25 ECU. If you assign 0 points to the last member of the group, it has no additional cost to you. For this example, the total cost of the points that you have assigned is 27 ECU (2 + 25 + 0). These costs are displayed on your screen. As long as you have not clicked on the [OK] button, you may modify your point assignment decision.

If you assign zero points to an individual, you do not change his earnings. However, if you distribute one point to an individual, you reduce his earnings by 10% of his first stage earnings. If you assign him two points, you reduce his earnings by 20%, etc… So, the number of points that you assign him defines how much you reduce his earnings below his first stage level.

❖**Your earnings at the end of the second stage** are calculated by the computer in the following manner:

If fewer than 10 points have been assigned to you in stage two:
Your earnings at end of second stage = (earnings at the end of 1st stage)*[(10-number of points received)/10] – cost of points you assigned in stage two

If 10 or more points have been assigned to you in stage two:
Your earnings at end of second stage = – cost of points you assigned in stage two

Note that in the earnings calculation, a maximum of 10 points received can count against your earnings. For example, if you have received a total of 3 points from the other members of your group, you earnings are reduced by 30% from the amount you had at the end of the first stage. If you received 4 points, your earnings from the 1st stage is reduced by 40%. If you have received 10 or more points, you earnings are reduced by 100% of your first stage earnings. If this happens, and you have assigned points in stage two, you lose money for the round. The amount you lose in that case equals the cost of the points that you assigned to the other members of your group.

Therefore your earnings at the end of the second stage can be negative, if the cost of the points you assign is greater than your earnings in the first stage. However, in general, you can always be sure to avoid losses if you make your decisions in a particular way.

❑ **At the beginning of the third stage**, your screen will indicate the total number of points that each of the other members of your group has assigned to each other member, including you. Afterward, **you have another opportunity to reduce or to leave unchanged the earnings of each of the other members of your group by assigning points. You can assign these points to demonstrate a level of disapproval (10 points for the most severe disapproval, 0**

points for the absence of disapproval). Each point assigned to an individual reduces the recipient's earnings by another 10% of his stage 1 earnings.

Similarly, your earnings can be modified if the other members of your group assign points to you.

You must decide on the number of points that you will assign to each of the three other members of your group to reduce their earnings or leave their earnings unchanged. You must enter a number of points for each subject between 0 and 10 inclusive. If you do not wish to reduce an individual's earnings, enter the number 0.

If you assign points, you incur a cost that depends on the number of points distributed to each subject. The greater the number of points you assign, the higher your costs are. Your total cost is equal to the sum of the three individual costs you pay to assign points to each of the other group members. The cost to you of the points you assign is identical to the cost of assigning the same number of points in stage 2.

At the end of stage three, your final earnings for each round are calculated by the computer in following manner

If a total of fewer than 10 points have been assigned to you in stages two and three:
Your earnings for the round = (earnings at the end of 1st stage)*[(10 – total number of points received in the second and third stages)/10] – cost of points you assigned in stages two and three

If a total of 10 or more points have been assigned to you in stages two and three:
Your earnings for the round = – cost of points you assigned in stages two and three

* * *

If you have any questions about what you have read, please raise your hand. An experimenter will come by and answer your question right away.
To make sure that you have understood the rules, please answer the following questions.

1. Suppose that each member of the group has 20 ECU at the beginning of the round. No member of the group (including you) contributes a single ECU to the project.
What are your earnings after the first stage? _____ ECU
What are the earnings of the other members of the group? _____ ECU

2. Suppose that each member of the group has 20 ECU at the beginning of the round. You contribute 20 ECU to the project. Each of the other group members also contributes 20 ECU to the project.
What are your earnings after the first stage? _____ ECU
What are the earnings of the other members of the group? _____ ECU

3. Suppose that each member of the group has 20 ECU at the beginning of the round. The other three members of the group assign a total of 30 ECU to the project.
What are your earnings if you contribute 0 ECU to the project? _____ ECU
What are your earnings if you contribute 15 ECU to the project? _____ ECU

4. Suppose that each member of the group has 20 ECU at the beginning of the round. You contribute 8 ECU to the project
What are your earnings if the other members of the group contribute a total of 7 ECU to the project? _____ ECU
What are your earnings if the other members of the group contribute a total of 22 ECU to the project?

5. Suppose that in the second stage of the game, you assign 9, 5, and 0 points to the three other members of your group. What is the total cost to you of the points you have assigned? _____ ECU.

6. What is the cost to you if you have assigned a total of 0 points? _____ ECU

7. By what percentage are your earnings from the first stage reduced if you receive a total of zero points from the other members of your group? _____ %

8. By what percentage are your earnings from the first stage reduced if you receive a total of 4 points form other members of your group? _____ %

9. By what percentage are your earnings from the first stage reduced if you receive a total of 15 points form other members of your group? _____ %